APPENDIX C-6
Web Server Transaction Log Analysis Methodology

APPENDIX C-6
Web Server Transaction Log Analysis Methodology

Table of Contents

APPENDIX C-6
WEB SERVER TRANSACTION LOG ANALYSIS METHODOLOGY

1.0. INTRODUCTION

The assessment and evaluation of electronic networks and network-based resources is increasing in scope and application (Bertot & McClure, 1996a, 1996b; McClure & Lopata, 1996). Web server transaction log file analysis is a network-based assessment technique that is particularly useful when performed in conjunction with other ongoing evaluation activities. The investigators designed an experimental data collection technique to analyze usage of GILS records on a Web server at the Environmental Protection Agency (EPA).

The intent of the Web server log analysis involved four purposes. One intent was to determine the overall Web site traffic including the location of users, the portions of the site accessed, and the number of document downloads. The second purpose was to determine the use of the Web site GILS directory traffic including the location of users, portions of the site accessed, and number of document downloads (both hits and accesses). The third purpose was to experiment with developing new log analysis techniques that go beyond domain, hit, and browser counts. The fourth purpose was to assist Federal agencies that operate Web-based GILS servers to develop, implement, and maintain ongoing log file analysis.

2.0. WEB SERVER TRANSACTION LOG ANALYSIS AS A RESEARCH TECHNIQUE IN SUPPORT OF THE PROJECT GOALS

Federal agencies make increasing use of the Web to provide access to Federal government information sources and in particular, to provide access to GILS records. In supporting GILS on the Web, Federal agencies have several important concerns that the investigators explored. These considerations include knowing what a server's traffic load is and the agency's overall ability to meet the demands of that traffic; knowing what a particular server's user community includes (e.g., accessing host IP address, browser, and operating system); knowing what users do while using the server; knowing both at what point and from where users accessed and left the server; and finally, knowing what problems users encountered during their server sessions. The investigators sought to develop a method by which Federal agencies could measure these indications of use to better manage their resources.

The investigators found that available log analysis software packages, commercially on the market, are generally inadequate to analyze log files in a variety of ways. The investigators reviewed multiple Web analysis software packages and analyzed them against four criteria: the ability to provide global and directory specific Web server analysis; the ability to distinguish between hits and accesses; the ability to determine user-specific actions though Web site session, and the ability to distinguish between unique and total referrals. None of the packages reviewed met all four criteria.

The investigators developed PERL-based scripts to analyze EPA log files that would provide all the required analysis capabilities. With these newly developed tools, the investigators created a mechanism that agencies could use to determine whether Web-based services are meeting the intended mission of the agency to provide public access to government information. By analyzing logs of user transactions, the investigators also attempted to asses a measure of user needs.

3.0. DESCRIPTION OF WEB SERVER TRANSACTION LOG ANALYSIS

The technique of Web server log analysis involved a three-fold process that included determining the types of information server administrators and decision makers need; developing a program that can parse through, manipulate, and present value-added information from the log files; and analyzing the information generated from the program. The investigators used four different log files which are automatically generated by Web servers (Rubin, 1996; Noonan, 1996; Novak and Hoffman, 1996). These four files are the access logs (e.g., hits), agent log

(e.g., browser, operating system), error log (e.g., download aborts), and referer logs (e.g., referring links). These log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a particular site.

Critical to understanding the type of data contained in these files is the distinction between a hit and an access. A hit is any file from a web site that a user downloads. If a user downloads a web page that has 6 images on it, then that user "hit" the web site seven times (6 images + 1 text page). An access (sometimes called a page hit) is an entire page downloaded by a user regardless of the number of images, sounds, or movies on the page. If a user downloads a web page that had 6 images on it, then that user just accessed one page of the web site.

One failing in most of the commercially available log analysis software packages is that the software counts the number of hits a server receives, rather than the number of accesses. The hit count reflects the number of items (e.g., images) downloaded when a user accesses a particular page. If a site has an image file, such as a Federal agency logo on multiple pages, that image will more than likely be the most frequently downloaded "hit" item on the site. Analysis information such as this is relatively useless in determining the site's actual usage.

4.0. TRANSACTION LOG ANALYSIS DATA COLLECTION AND DATA ANALYSIS ACTIVITIES

The investigators selected the EPA Web server from which to collect transaction log analysis files. Both the Department of Defense (DoD) and EPA offered to work with the investigators and the choice of agency became a decision of convenience and size. The size of the files at EPA ranged from 8 megabytes to 26 megabytes each on a daily basis, and this size was significantly smaller than the file sizes from DoD. In all, investigators used approximately 560 megabytes of log file data.

The investigators analyzed the access log, the agent log, the error log, and the referrer log. Log files were collected on February 8, 1997 and February 15, 1997. Each log file included a week's worth of transactions. The resulting output (Web log file analysis PERL scripts and log files) together consumed approximately 1 gigabit of storage.

The development and pretesting of the PERL scripts required considerable effort. The Syracuse University script development team required the equivalent of 240 man-hours developing the scripts. An additional 100 man-hours were required to pre-test the scripts using several different log files from different servers, including a test data set from the Federal agency GILS server. Running the scripts on the 14 day period of EPA log files and outputting the analysis into a usable format required an additional 100 man-hours. In total, therefore, the PERL script development process consumed approximately 420 man-hours.

The analysis of the EPA log files was performed on a Pentium 150 MHZ computer with 32 MB of RAM, and the analysis of each of the four daily log files took approximately 40 minutes.

5.0. LIMITATIONS OF TRANSACTION LOG ANALYSIS AS A RESEARCH METHOD

This method is exploratory and as such, is subject to further development. The use of this method encountered a few limitations.

The investigators had no guarantee that the files they received were complete data sets. There is a need to post the file size of the log files directly from the server such that agency staff responsible for this analysis can verify the file size against the downloaded files.

The investigators stored two weeks of log files from EPA as well as the PERL scripts. The resulting files took up nearly a gigabit of hard drive space. If an agency were to maintain this type of analysis on an ongoing basis, there is a need to dedicate a machine with adequate hard disk space for the task. It is also necessary to have a backup server or tape backup of the script and the log data files.

The investigators underestimated the number of daily referrals that the EPA server received. From analyzing the log analysis data, it is clear that a number of referrals came from search engines. The PERL scripts were not written to

extract this information. Future development of the scripts can help to determine not only what percentage of referrals come from search engines, but what search engines users tend to use and what search terms users enter.

In a more general context, it is important to interpret and consider log files as one component of a larger assessment activity for network services. While log files can provide Web administrators and others with critical server-related data, log files do not reflect user-based impact and outcome measures. As such, log files combine both user and technical perspectives on Web services.

The distinction between "hits" (downloads on an html page) and accesses (a downloaded html page) is critical. Software that counts only "hits" will not reflect the true nature of the site's use.
Agency use of commercially available transaction log analysis software may not readily support this distinction. Web administrators should not retrofit their log file analysis to the capabilities of this type of software.

Gaining access to and analyzing Web sever log files requires planning, coordination, and accountability. To engage in log file analysis activities, there needs to be a delegation of responsibility for making the files available (onsite or remotely), performing the analysis (onsite or remotely), interpreting the analyzed data, and reporting the findings. Moreover, such analysis needs to be performed and reported on an ongoing and regular basis.

There is a need to resolve these issues and move the ability to perform log file analysis forward. Log file data can provide user-based measures of Web-based resources if performed on a regular basis, incorporated into other electronic network assessment activities, and interpreted correctly.

6.0. PRIVACY ISSUES AND CONCLUSIONS

A major issue connected with data collection of HTTP transaction log analysis is privacy. In some cases, it is possible to trace directly back to a user, depending on the method of access a user has to a Web site. Web service providers need to develop policies as to how such data are to be used, if at all. This issue is particularly troublesome for public sector organizations, as such capabilities may violate privacy laws.

Web server transaction log analysis was an experimental data collect technique used by the investigators. This means of collecting data is relatively new and is one which will grow in importance to Federal agencies.